

Software Description

Natrix2 – Improved amplicon workflow with novel Oxford Nanopore Technologies support and enhancements in clustering, classification and taxonomic databases

Aman Deep¹, Dana Bludau¹, Marius Welzel², Sandra Clemens², Dominik Heider², Jens Boenigk^{1,3}, Daniela Beisser^{1,3}

¹ Department of Biodiversity, University of Duisburg-Essen, Universitätsstr. 5, 45141 Essen, Germany

² Faculty of Mathematics and Computer Science, University of Marburg, Hans-Meerwein-Str. 6, 35032 Marburg, Germany

³ Centre for Water and Environmental Research, University of Duisburg-Essen, Universitätsstr. 5, 45141 Essen, Germany

Corresponding author: Daniela Beisser (daniela.beisser@uni-due.de)

Abstract

Sequencing of amplified DNA is the first step towards the generation of Amplicon Sequence Variants (ASVs) or Operational Taxonomic Units (OTUs) for biodiversity assessment and comparative analyses of environmental communities and microbiomes. Notably, the rapid advancements in sequencing technologies have paved the way for the growing utilization of third-generation long-read approaches in recent years. These sequence data imply increasing read lengths, higher error rates, and altered sequencing chemistry. Likewise, methods for amplicon classification and reference databases have progressed, leading to the expansion of taxonomic application areas and higher classification accuracy. With Natrix, a user-friendly and reducible workflow solution, processing of prokaryotic and eukaryotic environmental Illumina sequences using 16S or 18S is possible. Here, we present an updated version of the pipeline, Natrix2, which incorporates VSEARCH as an alternative clustering method with better performance for 16S metabarcoding approaches and mothur for taxonomic classification on further databases, including PR², UNITE and SILVA. Additionally, Natrix2 includes the handling of Nanopore reads, which entails initial error correction and refinement of reads using Medaka and Racon to subsequently determine their taxonomic classification.

Key words: Amplicon sequencing, Amplicon Sequence Variants, community profiling, metabarcoding, microbiome, Operational Taxonomic Units, Snakemake workflow, ultra-long reads

Introduction

Analyzing nucleotide sequences of specific prokaryotic or eukaryotic DNA regions is the fundamental mechanism for advanced understanding of their biodiversity and biogeography. Amplicon sequencing of marker genes extracted from environmental samples can answer questions concerning presence, absence and even (relative) abundance of specific species or community composition. Due to constantly increasing demands, sequencing has developed rapidly in the recent decades. The cost and time intensive Sanger sequencing marks the beginning with



Academic editor: Thorsten Stoeck

Received: 12 July 2023

Accepted: 4 September 2023

Published: 24 October 2023

Citation: Deep A, Bludau D, Welzel M, Clemens S, Heider D, Boenigk J, Beisser D (2023) Natrix2 – Improved amplicon workflow with novel Oxford Nanopore Technologies support and enhancements in clustering, classification and taxonomic databases. Metabarcoding and Metagenomics 7: e109389. <https://doi.org/10.3897/mbmg.7.109389>

Copyright: © Aman Deep et al.

This is an open access article distributed under terms of the Creative Commons Attribution License (Attribution 4.0 International – CC BY 4.0).

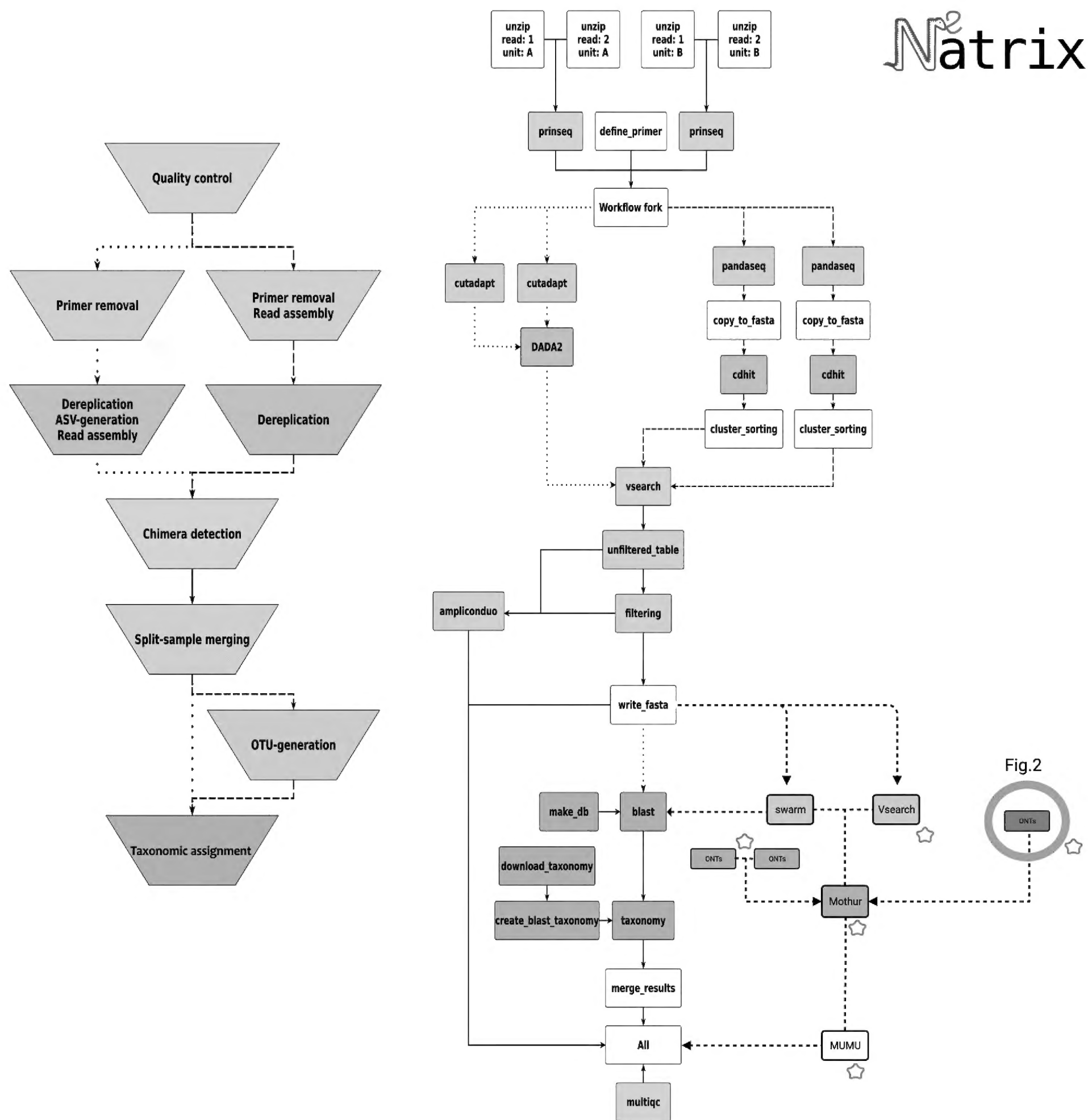


Figure 1. Schematic representation of the Natrix2 workflow. The processing of two split samples using AmpliconDuo is depicted. The color scheme represents the main steps, dashed lines outline the OTU and dotted edges outline the ASV variant of the workflow. Stars depict updates to the original Natrix workflow. Details on the ONT part are depicted in Fig. 2. (Created with BioRender.com).

further development to high-throughput sequencing like Illumina technologies to the latest real-time sequencing platform from Oxford Nanopore Technologies (ONT). Regardless of sequencing technology, raw sequencing reads need to be processed in multiple steps and clustered into taxonomically assigned sequence representatives for further analysis. Despite numerous available tools for each step, there are just few all-in-one and user-friendly workflows (Schloss et al. 2009; Callahan et al. 2016; Asbun et al. 2020; Tian and Imanian 2022).

For Illumina amplicon data, Natrix is one of few efficient workflows for read processing, OTU or ASV clustering and assigning amplicon sequencing reads

to taxonomy, with an adjustable workflow system (Welzel et al. 2020). It is an open-source pipeline that includes quality control, read assembly, dereplication, chimera detection, and taxonomic assessment. It utilizes Snakemake (Köster and Rahmann 2012) and bioconda (Grüning et al. 2018) for reproducibility and scalability. The pipeline executes various steps such as demultiplexing, adapter trimming, quality assessment with Cutadapt (Martin 2011), FastQC (Andrews 2010), MultiQC (Ewels et al. 2016), and PRINSEQ (Schmieder and Edwards 2011). PANDAseq (Masella et al. 2012) is used for primer defining and paired-read assembly. DADA2 (Callahan et al. 2016) can be used to generate ASVs. CD-HIT (Fu et al. 2012) performs dereplication in the OTU variant of the workflow. Chimeric sequences are detected using VSEARCH3 (Rognes et al. 2016) and split samples merged with AmpliconDuo (Lange et al. 2015). OTUs are generated using Swarm (v3) (Mahé et al. 2022). Finally, taxonomic assignments are identified using BLASTn (Altschul et al. 1990) against SILVA (Pruesse et al. 2007) or NCBI (Federhen 2012) databases. The final output comprises a comprehensive table with sequence information, abundances, and taxonomic data.

However, sequencing platforms undergoing a constant development, thus adaptations to new sequencing technologies are required. One of the latest technologies, Nanopore, is capable of producing read lengths of more than 800,000 base pairs (Jain et al. 2018), compared to Illumina reads with a maximum of 300 base pairs (Hu et al. 2021). However, its error rates are ranging from 6 to 8%, which is much higher than Illumina reads. Therefore, Nanopore data requires thorough processing to address these higher error rates. In addition to rapid advancements in sequencing platforms, classification methods have also evolved greatly in recent years. The constantly increasing number of reads produced per sequencing run and the associated computing capacity during processing, as well as the growth of gene reference libraries, have made this necessary (Ye et al. 2019). Whereas a few years ago the BLAST algorithm was the preferred classification tool for taxonomic assignment, nowadays classifiers with higher accuracy, lower computational capacity, and more specific reference databases are favored (Schloss et al. 2009; Gerlach and Stoye 2011; Wood and Salzberg 2014; Murali et al. 2018). The increasing number of microbial metabarcoding approaches has led to the development of databases specifically tailored to the research question. One of the many databases existing and already included in Natrix is SILVA, which is suitable for analysis of ribosomal subunit genes for prokaryotes and eukaryotes (Pruesse et al. 2007), while the NCBI database, which is likewise included, is suitable for a broad taxonomic classification of different species that do not necessarily belong to the same phylum (Federhen 2012). Instead of the often used ribosomal marker genes, the UNITE database uses the eukaryotic internal transcribed spacer (ITS) region located between two transcribed genes (Nilsson et al. 2019). Organismic groups including protists, fungi, metazoa or plants can be classified using databases such as PR². It contains nearly 200,000 sequences and annotations which are manually curated (Guillou et al. 2012). In addition to Swarm, we have also included VSEARCH clustering (Rognes et al. 2016) as an alternative to provide the user with more options and flexibility. It can be used as a drop-in replacement for Swarm in this existing workflow.

Natrix2, was thus extended to meet the above mentioned demands. On the one hand, it now includes specific pipeline options exclusively for Nanopore sequences. The automatic identification, reorientation and trimming of Nanopore

reads were integrated, as well as Naopore specific error correction and clustering. On the other hand, clustering and taxonomic classification was improved for Illumina sequences providing further clustering options and additional databases for other marker genes. General improvements include the restructuring of input and output files, error checking and a detailed description and how-to of a complete workflow including example sequences and configuration files on GitHub (<https://github.com/dbeisser/Natrix2>).

Package upgrade description

In the new version of Natrix, Natrix2, four major improvements have been integrated compared to the previous version (Fig. 1). i) the implementation of VSEARCH as an alternative clustering method, ii) the addition of mothur for taxonomic classification, iii) the extension to further databases and marker genes, and iv) the support of Nanopore sequence processing.

VSEARCH clustering

As an alternative to the already contained Swarm clustering algorithm (Mahé et al. 2022), VSEARCH (v2.15.2) was included for OTU generation by sequence similarity de novo clustering of Illumina reads, using a greedy heuristic clustering algorithm with a centroid approach (Rognes et al. 2016). The option for choosing the clustering algorithm was added to the configuration file. VSEARCH uses an adjustable sequence similarity threshold. By default it is set to 0.98, resulting in clustering of sequences into one OTU with a similarity of 98%. The integration of the optional VSEARCH clustering improves processing of prokaryotic sequences and expands the field of application for the Natrix pipeline. In order to enhance the accuracy and reliability of Operational Taxonomic Unit (OTU) generation from Illumina and Nanopore reads, the mumu post-clustering algorithm was implemented (<https://github.com/frederic-mahe/mumu>). Through the utilization of mumu, incorrect OTUs are effectively eliminated by considering both the sequence similarity and co-occurrence patterns of the reads, resulting in an improved representation of biodiversity.

Taxonomic classification and additional databases

In addition to BLAST searches used in the previous version of Natrix, the 'classify.seqs' function from the open-source mothur package was added to assign a taxonomy from a specific database defined in the configuration file (Schloss et al. 2009). Mothur provides packages and functions that are used for molecular analysis of community sequence data. Instead of creating alignments between sequenced reads and database references, mothur uses a kmer-based approach. Kmers are used to calculate the probability of sequences belonging to a specific taxonomy. Sequences with the highest probability will be assigned to the appropriate taxonomy. With the incorporation of the PR² and UNITE databases in addition to the SILVA and NCBI nr databases, new marker genes and organismic groups can now be addressed. The PR² (Protist Ribosomal Reference) database focuses on 18S rRNA metabarcoding approaches not only for protists, but also for fungi, metazoa and plants (Guillou et al. 2012).

Through the curation of experts, the PR² database is a reliable complement to the Natrix pipeline, making it usable for various research approaches. With the added UNITE database additional taxonomic analysis with focus on the eukaryotic nuclear ribosomal ITS region is now possible (Nilsson et al. 2019). The addition of UNITE offers more than one million fungal reference sequences, making Natrix an optimal tool for fungal metabarcoding. Taxonomic classification by mothur is made available for both Illumina and Nanopore reads.

Nanopore support

As the first version of Natrix was designed for Illumina sequencing reads only, support for processing of Nanopore long-reads was added (Fig. 2). Nanopore support can be activated within the configuration file and Nanopore reads in FASTQ format are used as the initial starting file. Sequencing adapters and primer sequences are identified by Pychopper (v2), a tool provided by ONT, using a combination of global and local alignments (<https://github.com/epi2me-labs/pychopper>). Reads are afterwards trimmed and oriented into forward direction. Pychopper is automatically installed using conda, and therefore version controlled. Next to its trimming and orienting options, Pychopper writes fused reads in an additional output file, from which reads are trimmed and orientated subsequently with a specific read rescue option. Afterwards, Nanopore reads are clustered and error corrected using CD-HIT (v4.8.1) (Li and Godzik 2006) for clustering and Medaka (v1.7.2) (<https://github.com/nanoporetech/Medaka>) and Racon (v1.4.13) (Vaser et al. 2017) for error correction. First, fasta transformed reads are clustered based on a similarity threshold algorithm and representatives are mapped against the initial fasta files with Minimap2 (v2.26) (Li 2018). Second, the initial fasta files, clustering and mapping data are used for the generation of consensus sequences of higher quality. Here, Racon is using a distance- and quality-based alignment algorithm, whereas Medaka is based on a neural network algorithm for creation of error corrected consensus sequences. Last, consensus sequences are again aligned by Minimap2 against the initial fasta files for identification of corresponding read numbers per consensus. Afterwards, the VSEARCH uchime3_denovo algorithm is still used for chimera removal of Nanopore sequences (Rognes et al. 2016) before the Nanopore reads are filtered and used further for taxonomic classification via BLAST or mothur (Altschul et al. 1990; Schloss et al. 2009).

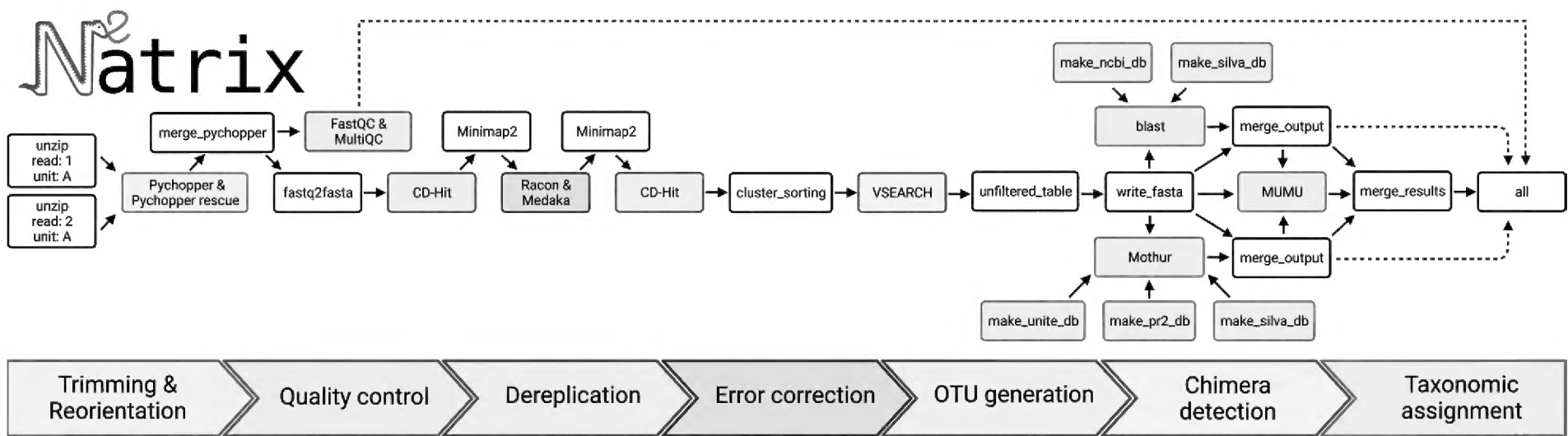


Figure 2. Schematic diagram of processing nanopore reads with Natrix2 for OTU generation and taxonomic assignment. The color scheme represents the main steps of this variant of the workflow. (created with BioRender.com).

Conclusion

With the upgraded version of Natrix, processing of Nanopore short and long sequencing reads, including orientation, trimming, clustering and error correction, is possible. In addition, Illumina and Nanopore reads can now be taxonomically assigned via mothur and the accuracy of OTU clustering is enhanced via mumu post-clustering. Optionally, VSEARCH can now be used for clustering Illumina reads. The implementation of PR² and UNITE as new databases makes Natrix2 a reliable tool for diverse metabarcoding approaches and now offers processing of sequences originating from other organismic groups like fungi, metazoa and plants or further marker genes like ITS.

Project description

Title: Natrix2 – Improved amplicon workflow with novel Oxford Nanopore Technologies support and enhancements in clustering, classification and taxonomic databases.

Study area description: Amplicon sequence analysis.

Download page: <https://github.com/dbeisser/Natrix2>.

Programming language: Snakemake, Python, R, Bash.

Licence: MIT Licence.

Acknowledgements

We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Funding

This study was performed as part of the Collaborative Research Center (CRC) RESIST and analyses were performed by Project A04 (AD and DBE), funded by the German Research Foundation (DFG) – CRC 1439/1; project number 426547801.

Author contributions

Conceptualization: MW, DH, JB, DBE. Formal analysis: SC, DBI, AD. Methodology: AD, DBI, SC, DBE. Supervision: JB, DBE. Validation: AD. Visualization: AD, DBI. Writing – original draft: AD, DBI, DBE. Writing – review and editing: DBI, DH, JB, AD, SC, MW, DBE.

Author ORCIDs

Aman Deep  <https://orcid.org/0000-0001-7321-864X>

Dana Bludau  <https://orcid.org/0009-0003-3982-3178>

Marius Welzel  <https://orcid.org/0000-0002-4946-2156>

Sandra Clemens  <https://orcid.org/0000-0002-9710-1152>

Dominik Heider  <https://orcid.org/0000-0002-3108-8311>

Jens Boenigk  <https://orcid.org/0000-0001-8858-8889>

Daniela Beisser  <https://orcid.org/0000-0002-0679-6631>

Data availability

All of the data that support the findings of this study are available in the main text.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215(3): 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- Andrews S (2010) FastQC: a quality control tool for high throughput sequence data.
- Asbun AA, Besseling MA, Balzano S, van Bleijswijk JDL, Witte HJ, Villanueva L, Engelman JC (2020) Cascabel: A scalable and versatile amplicon sequence data analysis pipeline delivering reproducible and documented results. *Frontiers in Genetics* 11: e489357. <https://doi.org/10.3389/fgene.2020.489357>
- Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13(7): 581–583. <https://doi.org/10.1038/nmeth.3869>
- Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32(19): 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
- Federhen S (2012) The NCBI Taxonomy database. *Nucleic Acids Research* 40(D1): D136–D143. <https://doi.org/10.1093/nar/gkr1178>
- Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23): 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Gerlach W, Stoye J (2011) Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Research* 39(14): e91. <https://doi.org/10.1093/nar/gkr225>
- Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J (2018) Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods* 15(7): 475–476. <https://doi.org/10.1038/s41592-018-0046-7>
- Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J, Del Campo J (2012) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research* 41(D1): D597–D604. <https://doi.org/10.1093/nar/gks1160>
- Hu T, Chitnis N, Monos D, Dinh A (2021) Next-generation sequencing technologies: An overview. *Human Immunology* 82(11): 801–811. <https://doi.org/10.1016/j.humimm.2021.02.012>
- Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O’Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* 36(4): 338–345. <https://doi.org/10.1038/nbt.4060>

- Köster J, Rahmann S (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* 28(19): 2520–2522. <https://doi.org/10.1093/bioinformatics/bts480>
- Lange A, Jost S, Heider D, Bock C, Budeus B, Schilling E, Strittmatter A, Boenigk J, Hoffmann D (2015) AmpliconDuo: A split-sample filtering protocol for high-throughput amplicon sequencing of microbial communities. *PLoS ONE* 10(11): e0141590. <https://doi.org/10.1371/journal.pone.0141590>
- Li H (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18): 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13): 1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>
- Mahé F, Czech L, Stamatakis A, Quince C, de Vargas C, Dunthorn M, Rognes T (2022) Swarm v3: Towards tera-scale amplicon clustering. *Bioinformatics* 38(1): 267–269. <https://doi.org/10.1093/bioinformatics/btab493>
- Martin M (2011) Cudadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* 17(1): 1–10. <https://doi.org/10.14806/ej.17.1.200>
- Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012) PANDAsseq: Paired-end assembler for illumina sequences. *BMC Bioinformatics* 13(1): 1–31. <https://doi.org/10.1186/1471-2105-13-31>
- Murali A, Bhargava A, Wright ES (2018) IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* 6(140): e140. <https://doi.org/10.1186/s40168-018-0521-5>
- Nilsson RH, Larsson KH, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Kõljalg U, Abarenkov K (2019) The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research* 47(D1): D259–D264. <https://doi.org/10.1093/nar/gky1022>
- Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO (2007) SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* 35(21): 7188–7196. <https://doi.org/10.1093/nar/gkm864>
- Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: A versatile open source tool for metagenomics. *PeerJ* 10: 1–22. <https://doi.org/10.7717/peerj.2584>
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, van Horn DJ, Weber CF (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75(23): 7537–7541. <https://doi.org/10.1128/AEM.01541-09>
- Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27(6): 863–864. <https://doi.org/10.1093/bioinformatics/btr026>
- Tian R, Imanian B (2022) ASAP 2: A pipeline and web server to analyze marker gene amplicon sequencing data automatically and consistently. *BMC Bioinformatics* 23(27): 27. <https://doi.org/10.1186/s12859-021-04555-0>
- Vaser R, Sovic I, Nagarajan N, Sikic M (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research* 27(5): 737–746. <https://doi.org/10.1101/gr.214270.116>
- Welzel M, Lange A, Heider D, Schwarz M, Freisleben B, Jensen M, Boenigk J, Beisser D (2020) Natrix: A Snakemake-based workflow for processing, clustering, and taxo-

- nomically assigning amplicon sequencing reads. BMC Bioinformatics 21(1): e526. <https://doi.org/10.1186/s12859-020-03852-4>
- Wood DE, Salzberg SL (2014) Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15(46): R46. <https://doi.org/10.1186/gb-2014-15-3-r46>
- Ye SH, Siddle KJ, Park DJ, Sabeti PC (2019) Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178(4): 779–794. <https://doi.org/10.1016/j.cell.2019.07.010>